Explainable AI

Procheta Sen
University of Liverpool, United Kingdom

December 13, 2022

Table of contents

Explaining Explainabile AI

- What is Explainability?
  - It is a framework which helps to interpret the decision or predictions made by an AI model (e.g. machine learning or deep learning model).
  - The target audience for the interpretation can be a lay man or a stake holder or people having domain expertise.
  - The motivation for explainability is to build socially responsible and trustworthy AI models.

Example Use Case Scenario for Explainability

- Here the target audience are common people who do not have IR expertise.
- The lay person can understand what words or phrases in the document were the reason for which the documents were in the top k rank corresponding to the query.

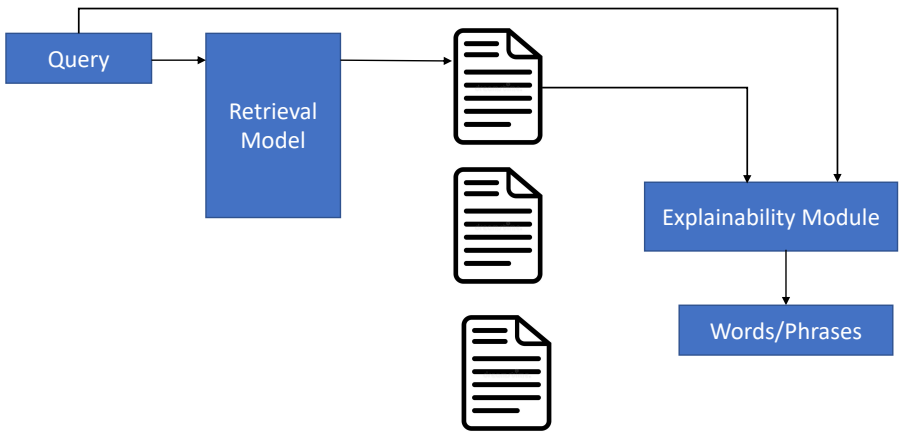


Figure: Example Explainer for an IR Model

Example Use Case Scenario for Explainability (Continued)

- The target audience are doctors. They have the domain expertise.

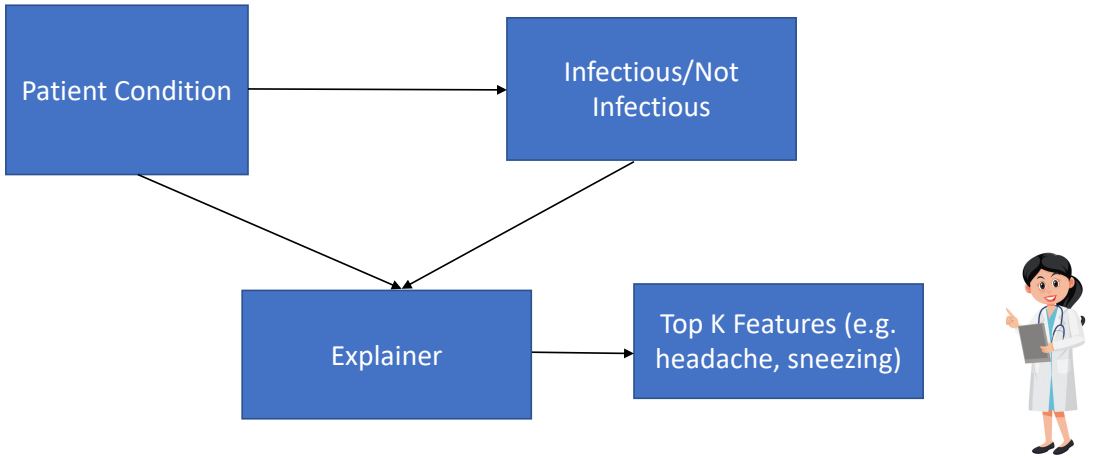


Figure: Explainer of an AI Model in Healthcare Domain

Example Scenario for Explainability (Use Case 3)

- Target audience are stakeholders.

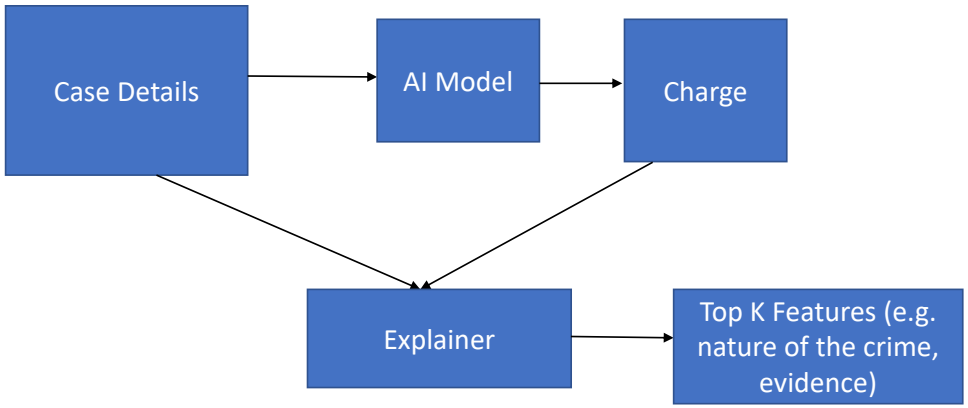


Figure: Explainer of an AI Model in Law Domain

Motivation for Explainability

What is the benefit of understanding an AI model?

- Is the model following non-expert intuition?
  - For the model shown in Use Case 1, if the model follows a lay man's intuition then they will have trust in the model.
- Is the model following an expert intuition?
  - For the model shown in Use Case 2, if the model prediction correlates with doctor's intuition then the model has learned well.

Motivation for Explainability (Continued)

- Helps the stakeholder
  - If the stakeholder has confidence in the model, he will be eager to invest in the project.
- Develop trustworthy model through recourse?
  - If the explanations are not trustworthy then the algorithm can take recourse and modify the model [Ross et al., 2021].
- Is the model fair?
  - Fairness is an important topic in today's world. The predictions made by an AI model should be fair in terms of gender race or any other socially sensitive attributes.
  - Explainability can help to understand whether the model is fair with respect to a attribute.

Explainability Research Trajectory[1]

- **2016-2019:** Posthoc Explanations [Ribeiro et al., 2016].
  - Research on building explainers for complex models.
- **2018-2022:** Critical Examination of SOTA [Slack et al., 2020].
  - Are explainability approaches trustworthy?
- **2020-2022:** Finding New Ways Forward [Lakkaraju et al., 2022].
  - Can we change the notion of explainability?

[1] Hima Lakkaraju, WiML, NeurIPS 2022 Talk

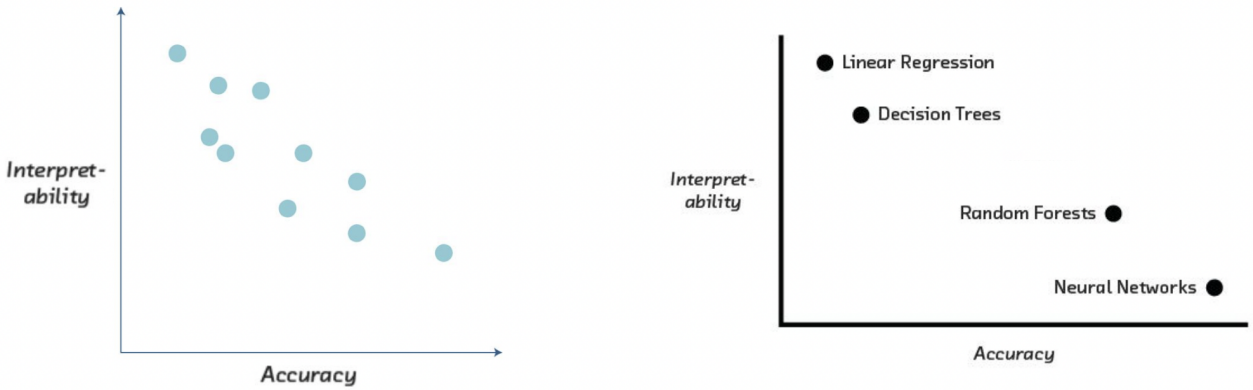
Properties of an Explanation [Lipton, 2018]

- Explainability is both important and Slippery.
- Transparency
  - Decomposability: Each part of the model (i.e. input, parameter, and calculation) admits an intuitive explanation.
  - Simulability: If a person can contemplate the entire model at once.
- Post-hoc Interpretability:
  - Presents a distinct approach to extracting information from learned models.
  - They often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning.

Explainability Vs. Accuracy

- In certain settings there may exist a tradeoff between explainability and accuracy [2].

Example



[2] Hima Lakkaraju, XAI Tutorial 2021

Approaches Towards Explainability

- Explaining without a separate explainer module.
  - The model takes as input human interpretable features.
  - The working principle of the model in inherently interpretable (e.g. Decision Tree).
- Explaining with a separate explainer module.
  - Use the model as a black box and try to explain the model with a much simpler architecture.
  - The explainer module tries to mimic the prediction of the original model.

Inherent Explanation Example: Decision Tree

- Decision tree checks different hierarchical conditions to eventually predict the output.

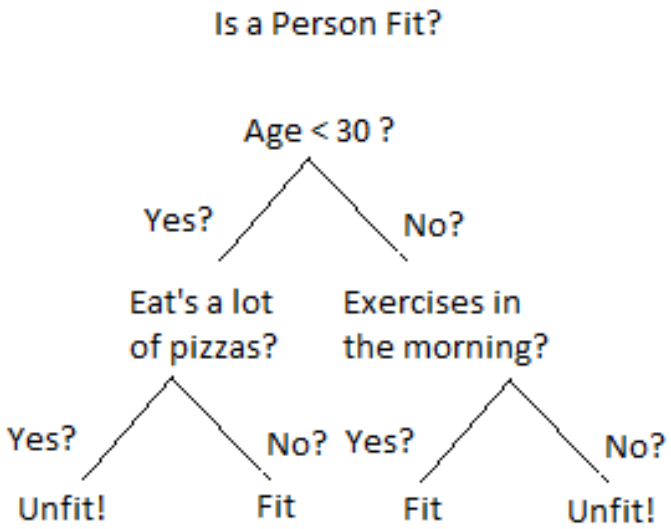


Figure: Example Decision Tree

Inherent Explanation Example in IR

- Model takes human interpretable features as input [Wang et al., 2013].
- SVM assigns weight to each feature.

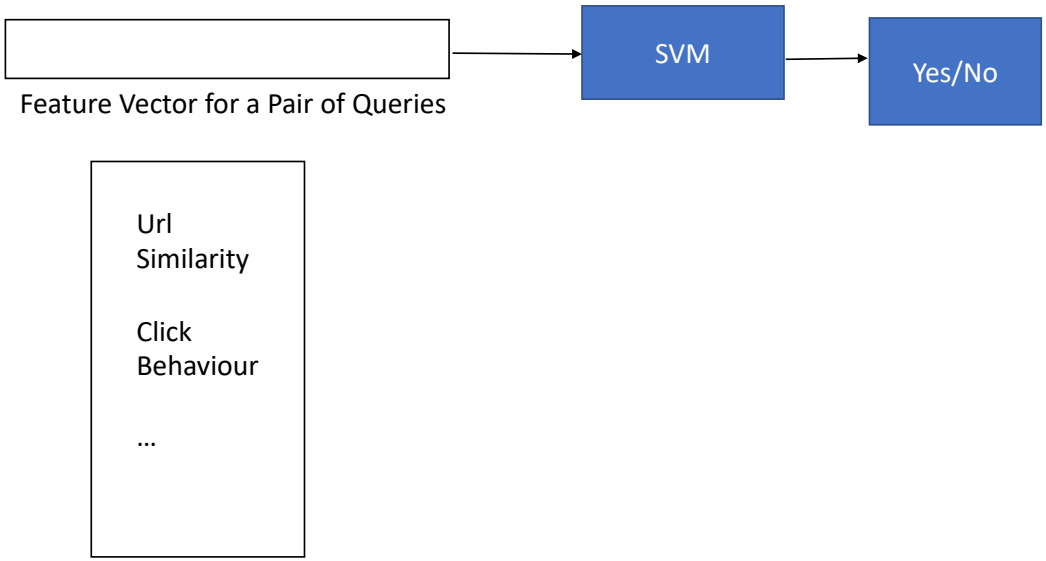


Figure: SVM Model for Predicting Label for a Pair of Queries.

Explaining with an Explainer

- Also known as post hoc explanations.
- Sometimes we only have access to a black box model.
- An inherently interpretable model is always preferable.

Explainability and Interpretability

- There is no formal definition of interpretability.
- Accroding to certain studies when the working principle of the model is explainable it is interpretable.
- Essentially inherently explainable models are known as interpretable models.

Types of Post-hoc Explanations

- Local Explanation
  - Takes a particular neighborhood around a sample point.

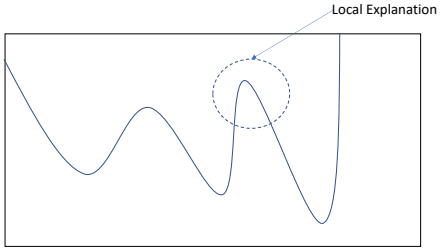


Figure: Local Explanation in a Particular Neighborhood.

  - More likely to give better explanations for complex boundary.
- Global Explanation
  - May be difficult for complex Models.
  - Gives an overview of the whole model.

Comparison of Local Vs. Global [3]

- **Local:** Explain individual predictions.
  Global: Explains complete behavior of the model.
- **Local:** Help unearth biases in the local neighborhood of a given instance
  Global: Help shed light on big picture biases affecting larger subgroups.
- **Local:** Help vet if individual predictions are being made for the right reasons.
  Global: Help vet if the model, at high level, is suitable for deployment.

[3] Hima Lakkaraju, XAI Tutorial AAAI 2021

Local Explanation Example in NLP

- Proposed a model [Bahdanau et al 2015] that jointly learns to align and translate.
- Align model estimates how well the inputs around position j and the output at position i match.
- α_i (i.e. the output of align model) are the local explanations.

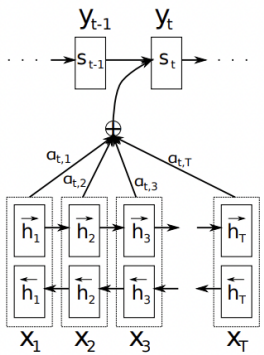
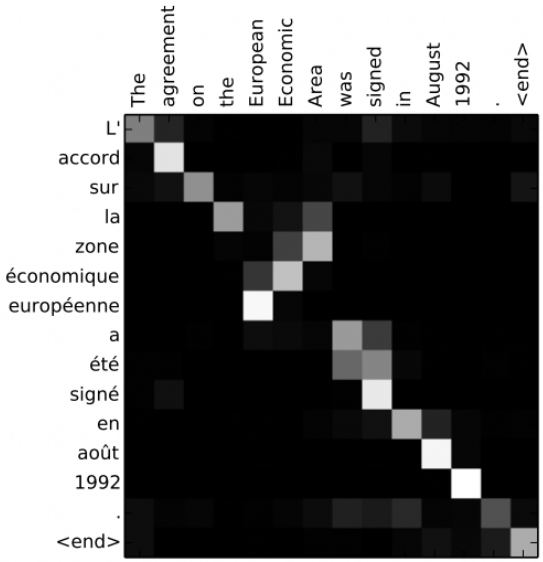


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.



Approaches Used for Local Explanations

- Feature Importance
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactual Examples

Feature Importance

- Train an explainer model (e.g. regression framework) based on human unterpretable input and model predicted output .
- The explainer will estimate the importance of different human interpretable features.
- Mostly model agnostic.

LIME: Sparse Linear Explanation [Ribeiro et al., 2016]

- Locally interpretable feature importance approach.
- Takes epsilon neighborhood around each point.
- Train a classifier for those points.
- The classifier estimates the importance of different features.

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

Figure: Sparse Linear Explanation Algorithm

LIME Global Explanation Algorithm [Ribeiro et al., 2016]

- Take average of weights across all local explanations.
- Choose the feature set that gives maximum coverage.

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B
  for all $x_i \in X$ **do**
    $\mathcal{W}_i \leftarrow$ **explain**(x_i, x_i') ▷ Using Algorithm 1
  end for
  for $j \in \{1 \dots d'\}$ **do**
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ **do** ▷ Greedy optimization of Eq (4)
    $V \leftarrow V \cup \text{argmax}_i\, c(V \cup \{i\}, \mathcal{W}, I)$
  end while
  return V

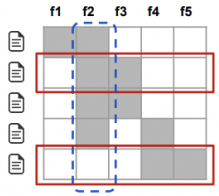


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Figure: Sparse Linear Global Explanation Algorithm

LIME Application Example

- Classifier predicts whether a document is about 'Christianity' or 'Atheism'.
- Figure shows important words for different algorithms.

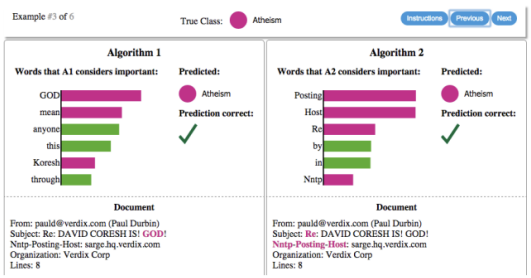


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

SHAP

- Another local explanation approach: SHAP (SHapley Additive exPlanation) Values.
- Uses a game theoretic approach for generating explanations.
- Compute shapley values of a conditional expectation function of the original model $E(f(z)|x_1)$.

SHAP (Continued)

- SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature.
- They explain how to get from the base value E[f(z)] that would be predicted if we did not know any features to the current output f(x).

SHAP Application in NLP

- The increment/decrement brought by each word is shown in the plot. For example, the word "amazing", with its TF-IDF score of 0.086, contributed to a 0.06 increment of the prediction.

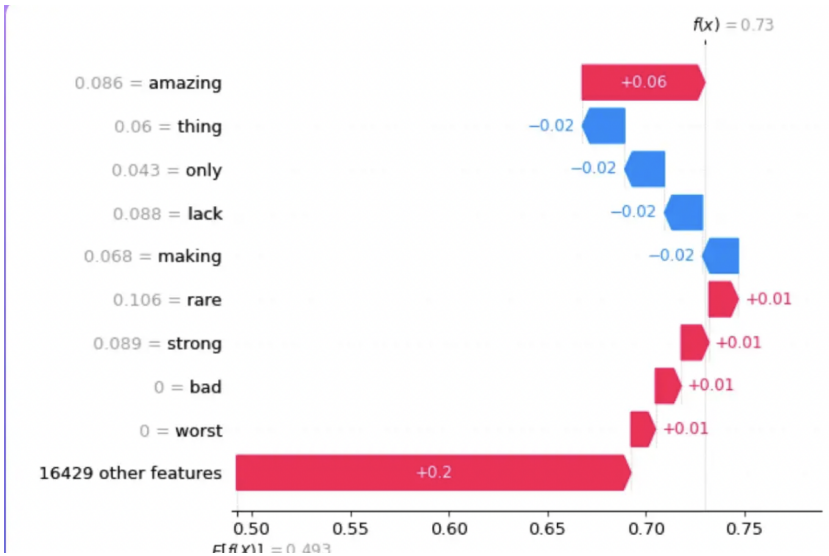


Figure: Shap Values for a Movie Review Classifier.

Example Application of LIME in IR (LIRME)

- Create different variations of a document D using different sampling approaches.
-
- Explain a document wrt a query and a retrieval model.
- Objective Function

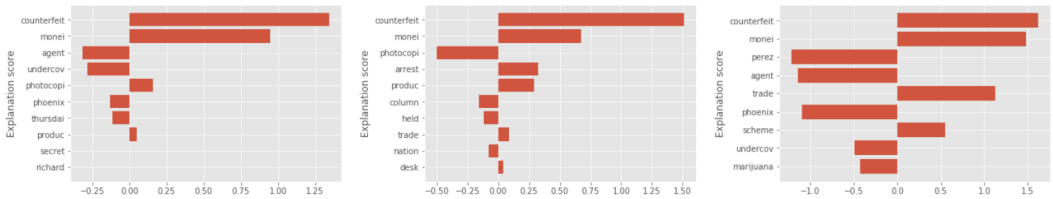
$$L(D, Q, \sigma, \theta) = \sum_{i=1}^{M} \rho(D, D_i')(S(D, Q) - \sum_{j=1}^{p} \theta_j w(t_j, D_i'))^2 \quad (1)$$

- θ_j is a p dimensional vector showing the importance of a term t in $S(D, Q)$.

LIRME (Document Sampling Approaches)

- **Uniform Sampling:** Sample terms with a uniform likelihood (with replacement).
- **Biased Sampling:** Set the sampling probability of a term proportional to its tf-idf weight seeking to generate sub-samples with informative terms.
- **Masked Sampling:** Specify a segment size, say k, and then segment a document D into $|D|K|$ number of chunks. Each subsample can comprise a set of chunks.

Example Explanation Diagram



(a) Uniform sampling (b) Tf-idf sampling (c) Masked samples ($v = 0.1$, $k = 5$)

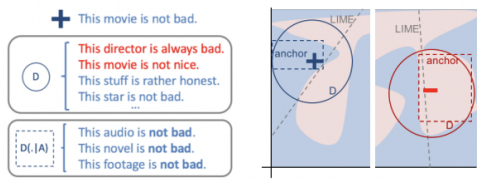
Figure 4: Visualization of explanation vectors $\hat{\Theta}(Q, D)$ estimated for a sample (relevant) document 'LA071389-0111' (D) and query (Q) 'counterfeiting money' (TREC-8 id 425). The Y-axis shows explanation terms, while the X-axis plots their weights.

Figure: Explanation Output From LIRME

LIRME: Evaluation Approaches

- **Explanation Consistency:** A particular choice of samples around the pivot document, D, should not result in considerable differences in the predicted explanation vector. Computes correlation between predicted and ground truth ranking of terms.
- **Explanation Correctness** Computes similarity between explanation vector terms $\theta(Q, D)$ and relevant terms $R(Q)$.

Rule Based Explanation



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$ (b) Two toy visualizations

Figure 2: Concrete example of $\mathcal{D}$ in (a) and intuition (b)

Instance	If	Predict
I want to play(V) ball.	previous word is PARTICLE	play is VERB.
I went to a play(N) yesterday.	previous word is DETERMINER	play is NOUN.
I play(V) ball on Mondays.	previous word is PRONOUN	play is VERB.

Table 1: Anchors for Part-of-Speech tag for the word "play"

- Local decision boundary may not be linear.
- Identifies the condition under which the classifier has the same prediction.
- The sufficient condition is known as anchor conditions.

Saliency Maps

- What part of the input is important for output?
- Depends on the gradient.
- It has been widely used for image classification problems.
- Also known as heatmap or feature attribution approaches.
- Varaints: Input-Gradient, Smooth-Gradient, Integrated gradient

Explanability Beyond Classification: Large Language Models

- Provides Contrastive Explanations for Large Language Models.
- Explainer looks for salient input tokens that explain why the model predicted one token instead of another.
- Contrastive explanations are better than non-contrastive explanations in verifying grammatical phenomena.

Explaining Large Language Models

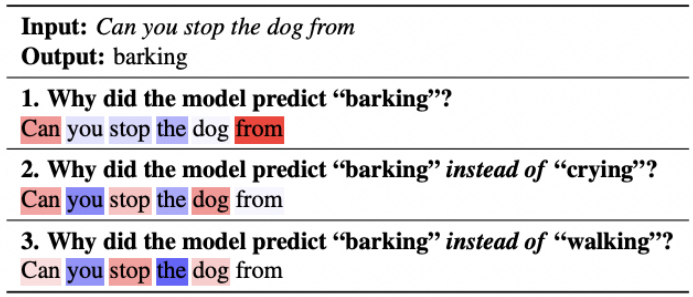


Figure: Explainer for Large Language Models

Explanability beyond classification: Large Language Models

- Focus on explanations that communicate why a computational model made a certain prediction.
- Compute saliency scores $S(xi)$ over input features xi to reveal which input tokens are most relevant for a prediction.
- The higher the saliency score, the more xi supposedly contributed to the model output.
- Contrastive explanations attempt to explain why given an input x the model predicts a target yt instead of a foil yf .
- Saliency score

$$g(x_i) = \Delta_{x_i} q(y_t|x) \tag{2}$$

Prototype based Approach [4]

- Explain a model with natural or synthetic example.
- What kind of models it is going to misclassify ?
- Which input maximally influence an input?

[4] Hima Lakkaraju, XAI Tutorial AAAI 2021

Prototype based Explanation via Influence Functions [5]

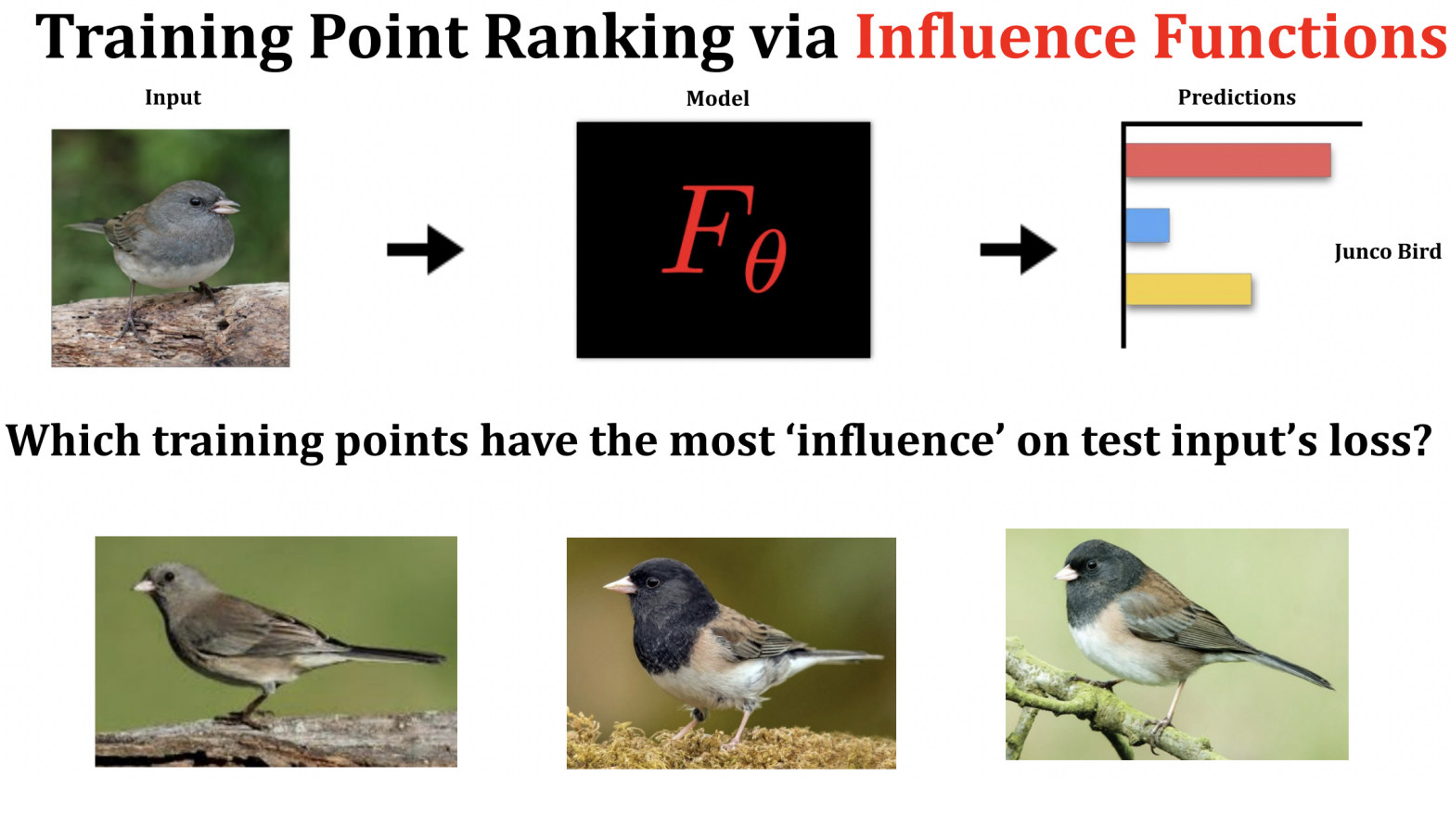


Figure: Junco Bird Classification problem Explanation

[5] Hima Lakkaraju, XAI Tutorial 2021

Influence Function [6]

- A statistical tool used in robust statistics for assessing the effect of a sample in regression parameter.
- Equation describing the influence of a sample on test input loss.

$$I_{z_j, z_{test}, \theta} = I(z_t est, \theta)^T H_\theta \delta_\theta I(z_j, \theta) \quad (3)$$

- Challenges: Computing hessian vector product can be tedious.

[6] Hima Lakkaraju, XAI Tutorial 2021

Counterfactual Explanations [7]

- What features need to be changed?
- How much to flip a model's prediction?

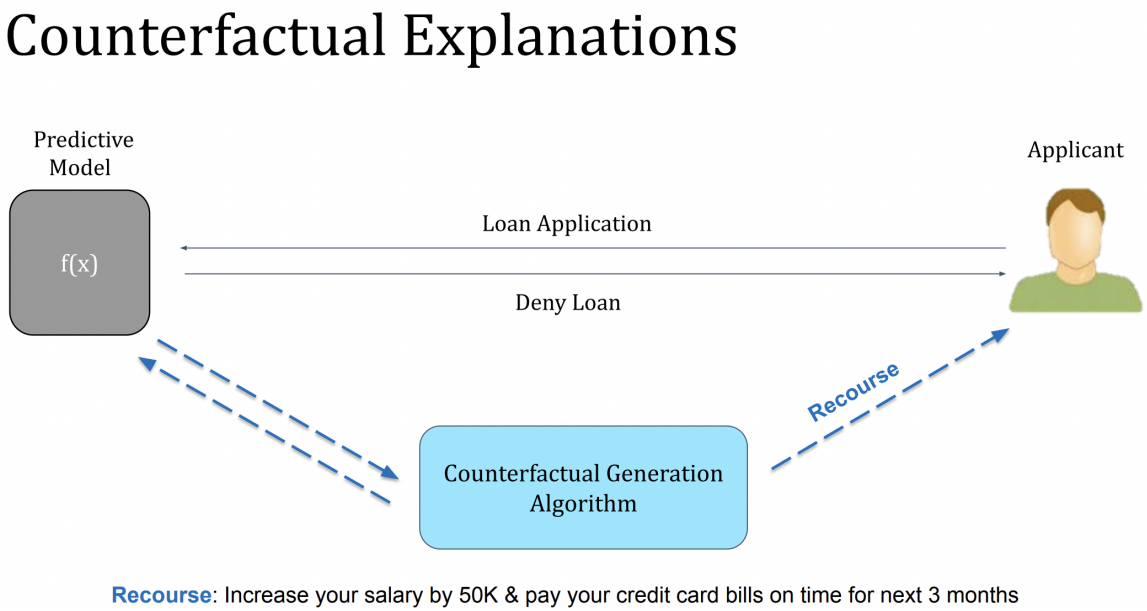


Figure: Counterfactual Explanation in a Loan Prediction Algorithm

[7] Hima Lakkaraju, XAI Tutorial 2021

Counterfactual Explanations (Continued)

Approaches for Counterfactual Explanations.

- Minimum Cost Counterfactual.
- Feasible and minimum cost.
- Causally feasible counterfactual.

Approaches for Global Explanations

- Collection of local explanations
- Model Distillation
- Representation Based

Explanation within a Ranking Model

- For each retrieval model and for each query train a regression classifier based on the fundamental features.
- Choose randomly K number of queries for a particular model.
- The contribution of each feature in a particular retrieval model is the average weights learned across K queries.

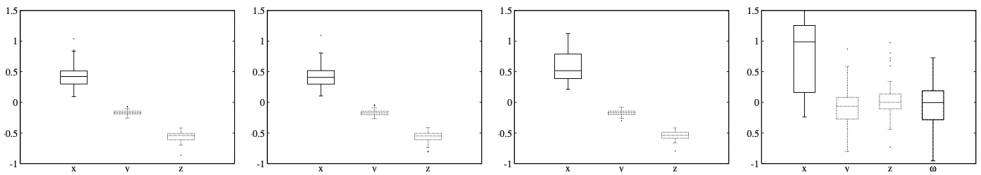


Figure 1: Box-plot of parameter vectors θ for BM25, LM-JM, LM-Dir and DRMM (in order from left-right).

Figure: Explanation Output From LIRME

Explaining Document Scores within and across Ranking Models

- The explanation units are the fundamental features of a retrieval model.
- Features Considered: Term Frequency, Document Length, Collection Frequency, Semantic Similarity
- This is a global explanation.

Explaining within a Ranking Model

Why does a model M retrieve a document D_1 at rank r_1 and D_2 at r_2 ($r_2 > r_1$ without loss of generality) for a query Q?

- For a particular feature the contribution is defined as follows

$$\bar{\phi}_x(D) = \frac{1}{|Q \cap D|} \sum_{w \in |Q \cap D|} \phi_x(w, D), \tag{4}$$

- Relative Drop Measure

$$\Delta_{x,D,D_{top}} = \frac{\bar{\phi}_x(D_{top}) - \bar{\phi}_x(D)}{\bar{\phi}_x(D_{top})}, \tag{5}$$

- Fidelity Score Computation

$$\xi(D, D_{top}) = (\Delta_{x,D,D_{top}}, \Delta_{y,D,D_{top}}, \Delta_{z,D,D_{top}}) \cdot \vec{\theta}. \tag{6}$$

Explaining Within a Ranking Model

The cases where the fidelity scores, $\xi_\gamma(M_1, M_2) > 0, \gamma = \{x, y, z\}$, and the matched values of the corresponding feature components (e.g. x for term frequency) act as the plausible *explanation*

Fidelity Score Illustration

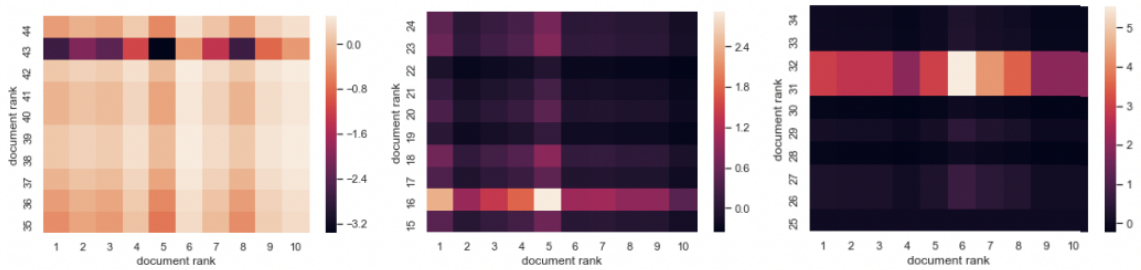


Figure 2: ξ_x, ξ_y and ξ_z distributions (left-right) for different rank pairs in BM25.

Figure: Fidelity Scores for BM25

Fidelity Score Computation

Why does a model M_1 retrieve a document D at position r_1, whereas model M_2 retrieves D at r_2 for a query Q?

- $$\Delta_s(D, M_1, M_2) = \delta_s(Q, D, M_2) - \delta_s(Q, D, M_1), \text{ where}$$
$$\delta_s(Q, D, M) = \frac{s(Q, D_{top}, M) - s(Q, D, M)}{s(Q, D_{top}, M)}. \quad (7)$$

- $$\xi(M_1, M_2) = \Delta_s(D, M_1, M_2) \cdot \Delta(M_1, M_2), \text{ where}$$
$$\Delta(M_1, M_2) = \vec{\theta}(M_1, Q) - \vec{\theta}(M_2, Q) \quad (8)$$

Model Distillation for Global Explanations

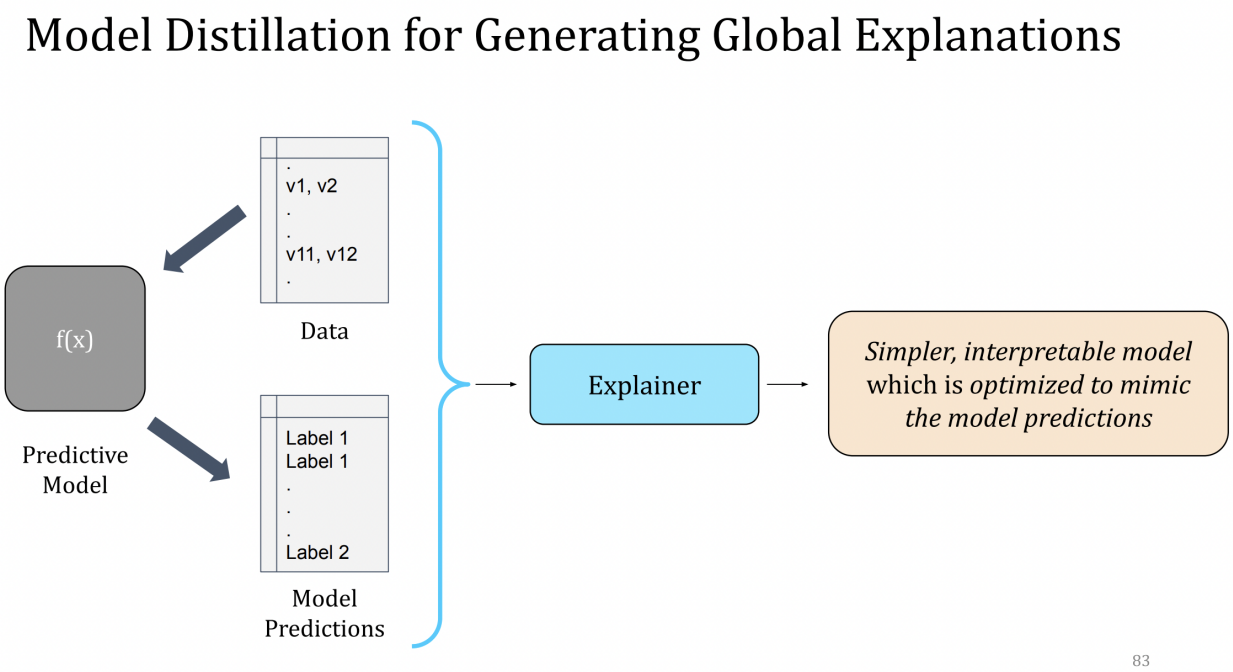


Figure: Explanation Using Model Distillation

Model Distillation for Global Explanations: Example Approach

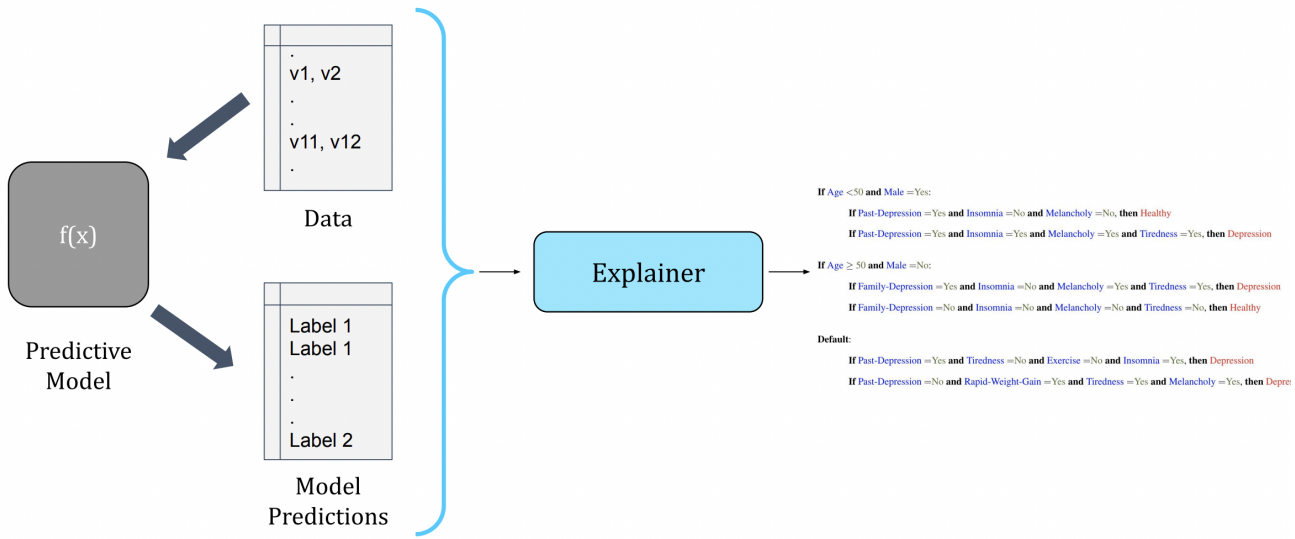


Figure: Explanation Using Model Distillation

Representation Based Approaches

- Estimates the importance of intermediate layers in a deep neural network model.
- Identify the human labelled concepts.
- Estimates the response of hidden variables to these concepts.
- Quantify the alignment of hidden variable-concept pairs.
- Some of the notable approaches include Network Dissection and TCAV.

Explaining neural network embedding in NLP [Liu et al 2018]

- A posthoc explanation approach to interpret network embeddings.
- Explanation is presented in terms of taxonomy.

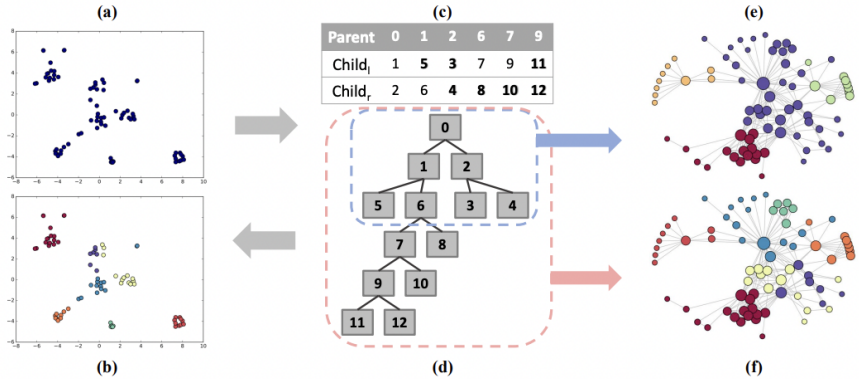


Figure 2: A taxonomy extracted by hierarchical clustering on embeddings from node2vec [22] on the Les-Misérables network. (a): Visualization of original embedding results. (b): Visualization of embeddings after clustering where $C = 7$. (c): Taxonomy represented in a table, where bold numbers denote leaves. (d): Taxomony represented by a tree. (e): Visualization of the network with 4 clusters obtained from a subtree. (f): Visualization of the network with 7 clusters obtained from the taxonomy, which corresponds to the embedding visualization in figure (b).

Figure: Taxonomy Explanation Example in NLP

Evaluation of Explanations [8]

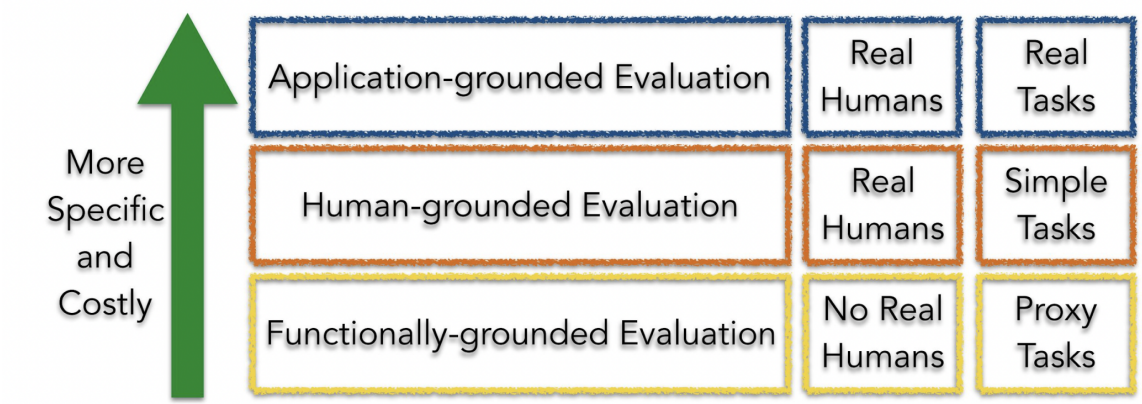


Figure: Categorization of Dofferent Evaluation Approaches

[8] Hima Lakkaraju, XAI Tutorial 2021

Understanding the Behaviour [9]

- Delete a feature and check the model's performance

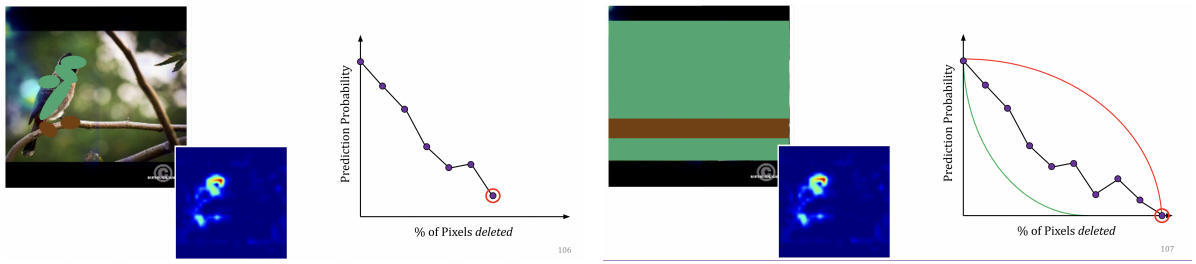


Figure: Classifier Performance at different levels of Deletion of Features

[9] Hima Lakkaraju, XAI Tutorial 2021

Predicting User Behaviour [10]

- Check whether a user can learn about the model behaviour from explanations.

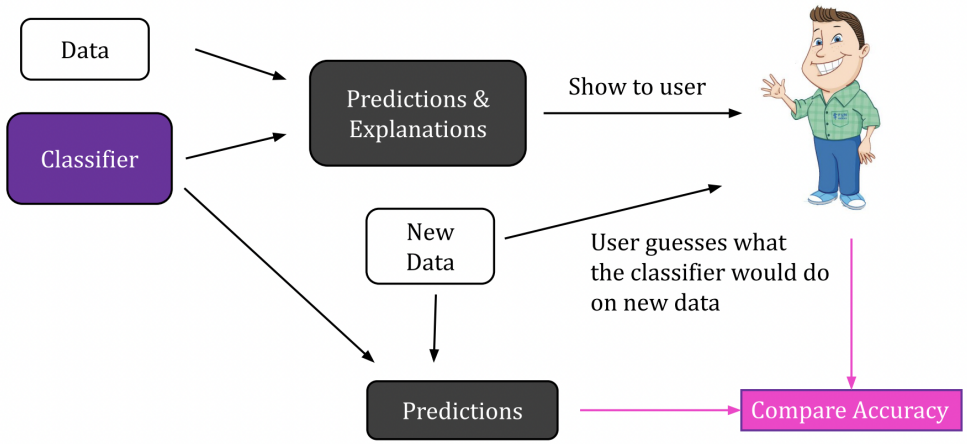


Figure: Predicting User Behavior Evaluation Architecture

[10] Hima Lakkaraju, XAI Tutorial 2021

Ask a Human [11]

- Do you trust
- Could they compare between classifiers?

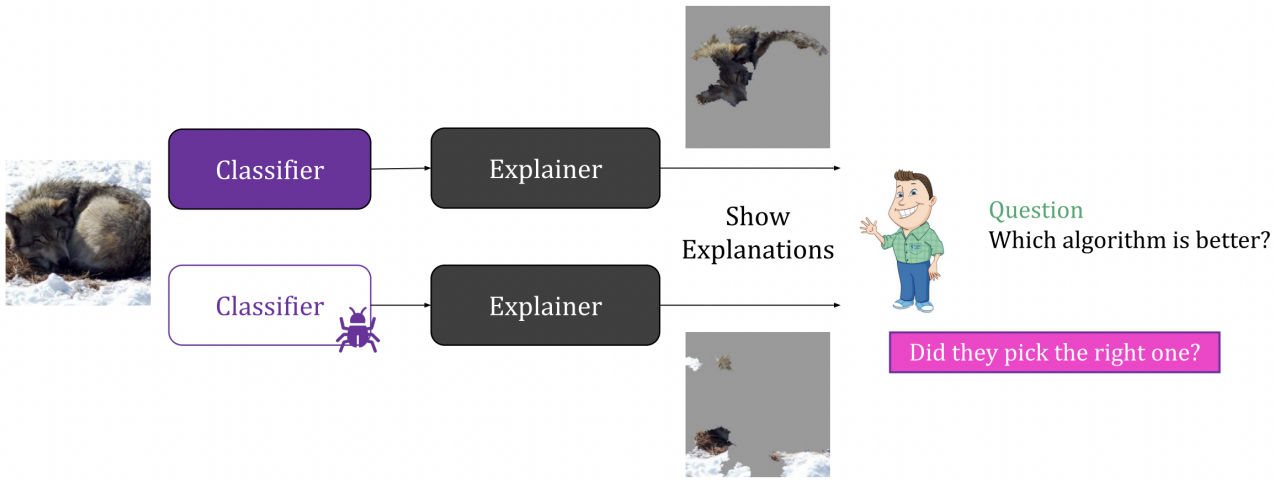


Figure: Explanation Evaluation by asking a Human

[11] Hima Lakkaraju, XAI Tutorial 2021

Limitations of Evaluation for Explainability [12]

Limitations of Evaluating Explanations

- Evaluation setup is often very easy/simple (or unrealistic)
  - E.g. "bugs" are obvious artifacts, classifiers are different from each other
  - Instances/perturbations create out-of-domain points
- Sometimes flawed
  - E.g. is model explanation same as human explanation?
- Automated metrics can be *optimized*
- User studies are not consistent
  - Affected by choice of: UI, phrasing, visualization, population, incentives, …
  - ML researchers are not trained for this ☹
- Conclusions are difficult to generalize

[12] Hima Lakkaraju, XAI Tutorial 2021

Limitations of Post Hoc Explainability [13]

- **Faithfulness/Fidelity**
  - Some explanation methods do not '*reflect*' the underlying model.
- **Fragility**
  - Post-hoc explanations can be easily manipulated.
- **Stability**
  - Slight changes to inputs can cause large changes in explanations.
- **Useful in practice?**
  - Unclear if a data scientist (ML engineer)/end-user can use explanations to isolate errors, improve 'trust' or simulate the model.

13 14

Adverserial Attacks on Explainability

- Modify input with small perturbation.
- Model output won't change.
- Observe the change in explanation under adversarial attack.
- Commonly known attacks: 'shift Attack, Augmented Loss Function Attack, Passive and Active Fooling Loss Augmentation Attack.
- LIME, SHAP suffer from these types of attacks.
- Some defense algorithms developed to handle the adversarial attacks are Hyperplane method, Autoencoder approach to prevent expalanation manipulation.

Simultaneous Training of an Explainer and the Model

- Propose an information-based framework for instancewise feature selection.
- Given a feature vector the model attempts to create all possible subset of the feature vector.
- For each variant the model computes the mutual information between the response variable and input.
- Higher mutual information denotes higher dependency between the output and the input feature subset.
- Eventually the explainer learns the importance of different features in the model.

L2X: Learning to Explain

Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.
negative	negative	You get 5 writers together, have each write a different story with a different genre, and then you try to make one movie out of it. Its action, its adventure, its sci-fi, its western, its a mess. Sorry, but this movie absolutely stinks. 4.5 is giving it an awefully high rating. That said, its movies like this that make me think I could write movies, and I can barely write.
negative	positive	This movie is not the same as the 1954 version with Judy garland and James mason, and that is a shame because the 1954 version is, in my opinion, much better. I am not denying Barbra Streisand's talent at all. She is a good actress and brilliant singer. I am not acquainted with Kris Kristofferson's other work and therefore I can't pass judgment on it. However, this movie leaves much to be desired. It is paced slowly, it has gratuitous nudity and foul language, and can be very difficult to sit through. However, I am not a big fan of rock music, so its only natural that I would like the judy garland version better. See the 1976 film with Barbra and Kris, and judge for yourself.
positive	negative	The first time you see the second renaissance it may look boring. Look at it at least twice and definitely watch part 2. it will change your view of the matrix. Are the human people the ones who started the war? Is ai a bad thing?

Table 3. True labels and labels from the model are shown in the first two columns. Key sentences picked by L2X highlighted in yellow.

Figure: L2X output for a Classifier

Simultaneous Training of an Explainer and the Model [Nguyen and Rudra, 2022]

- Proposed a rationale aware contrastive learning based approach to classify and summarize crisis related microblogs.
- In the first stage the model learns to extract rationales using a multi-task architecture.
- In the second stage the extracted rationales along with tweets are used for classification.

Figure 1: Examples of tweets. Tweet labels are in bold, and rationale snippets are in blue.

Figure: L1 distance between a pair of explanations

Comment on Explanability in NLP

- Classification problems can be explained with LIME like architecture in general.
- Counterfactual explanations are useful for text generation models.

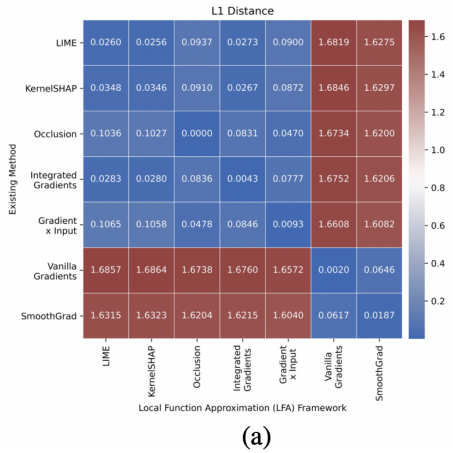
Comment on Explanability in IR

- Can be explained based on query-document pair. These are local explanations.
- Can be explained based on the features like term frequency, document frequency. This is a global explanation.
- Relevance and non-relevance class can be explained separately.
- Counterfactual explanation is more suitable for non-relevant documents.

Which Explanation Should I Choose?

- Proposed a linear function approximation (LFA) framework to formlate any explanabilty algorithm.
- Popular explanation algorithms like LIME follows LFA.
- Proposed 'no free lunch theorem' for explanation methods which demonstrates that no single explanation method can perform local function approximation faithfully across all neighbourhoods.
- The disagreement occurs because different methods approximate the black box model over different neighborhoods using different loss functions.
- If the explanation model perfectly approximates the underlying model when it is able to do so, then it should be chosen for explanation.

Which Explanation Should I Choose?



(a)

Figure: L1 distance between a pair of explanations

- For continuous data, use additive continuous noise methods (e.g. SmoothGrad, Vanilla Gradients).
- For binary data, use binary noise methods (e.g. LIME).
- Within each domain, choosing among appropriate methods boils down to determining the perturbation neighbourhood

Explanability Beyond Classification: RL [15]

[Madumal et. al., 2019]

Beyond Classification: Explainability for RL

- Causal explanations of the behavior of model free RL agents
- Generate explanations of agent behaviour based on counterfactual analysis of the causal model

Explaining the actions of a StarCraft II agent

Question	Why not *build_barracks* (A_b)?
Explanation	Because it is more desirable to do action build_supply_depot (A_s) to have more Supply Depots (S) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

[15] Hima Lakkaraju, XAI Tutorial 2021

Explanability Beyond Classification: RL [16]

[Madumal et. al., 2019]

Beyond Classification: Explainability for RL

- Causal explanations of the behavior of model free RL agents
- Generate explanations of agent behaviour based on counterfactual analysis of the causal model

Explaining the actions of a StarCraft II agent

Question	Why not *build_barracks* (A_b)?
Explanation	Because it is more desirable to do action build_supply_depot (A_s) to have more Supply Depots (S) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

[16] Hima Lakkaraju, XAI Tutorial 2021

Explanability Beyond Classification: RL [17]

- Model distillation using soft decision trees.
- Summarize agent behavior by identifying important states in a policy.

[17] Hima Lakkaraju, XAI Tutorial 2021

IR Explainability Survey Paper [Anand et al., 2022]

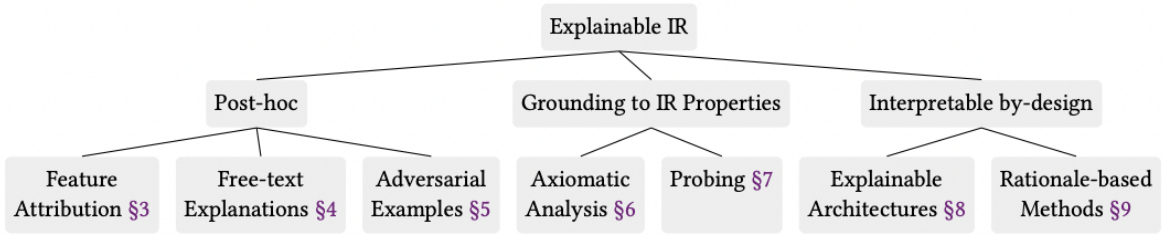


Figure: Explainer for Large Language Models

Scope for Future: TalkToModel

Componenents of the Model

- A natural language interface for engaging in conversations, making ML model explainability highly accessible.
- a dialogue engine that adapts to any tabular model and dataset, interprets natural language, maps it to appropriate explanations, and generates text responses,
- an execution component that constructs the explanations.

Scope for future

https://github.com/dylan-slack/TalkToModel

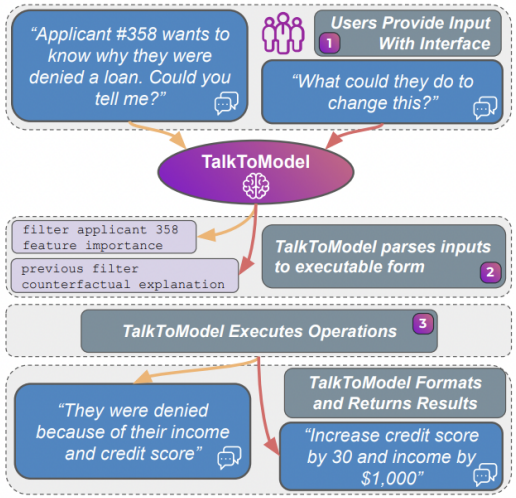


Figure 1: **An overview of TalkToModel:** Instead of writing code or using a dashboard, users engage in dynamic *open ended dialogues* with TalkToModel to understand models. First **(1)**, users supply natural language inputs using the interface. Next **(2)**, the dialogue engine parses the input into an executable representation of the question. After, **(3)**, the execution engine runs the operations. Finally, the dialogue engine formats the results into a response and provides it to the user.

Figure: Explainer for Large Language Models

References


Lakkaraju, H., Slack, D., Chen, Y., Tan, C., and Singh, S. (2022).
Rethinking explainability as a dialogue: A practitioner's perspective.
CoRR, abs/2202.01875.


Lipton, Z. C. (2018).
The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.


Nguyen, T. H. and Rudra, K. (2022).
Rationale aware contrastive learning based approach to classify and summarize crisis-related microblogs.
CIKM '22, page 1552–1562, New York, NY, USA. Association for Computing Machinery.